The integrated microbial genomes (IMG) system in 2007: data content and analysis tool extensions

Victor M. Markowitz[1], Ernest Szeto[1], Krishna Palaniappan[1], Yuri Grechkin[1], Ken Chu[1], I-Min A. Chen[1], Inna Dubchak[2], Iain Anderson[3], Athanasios Lykidis[3], Konstantinos Mavromatis[3], Natalia N. Ivanova[3] and Nikos C. Kyrpides[3]

[1]Biological Data Management and Technology Center, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, USA, [2]Genomics Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, USA, [3]Genome Biology Program, Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, USA

ABSTRACT

The Integrated Microbial Genomes (IMG) system is a data management, analysis and annotation platform for all publicly available genomes. IMG contains both draft and complete JGI microbial genomes integrated with all other publicly available genomes from all three domains of life, together with a large number of plasmids and viruses. IMG provides tools and viewers for analyzing and annotating genomes, genes and functions, individually or in a comparative context. Since its first release in 2005, IMG's data content and analytical capabilities have been constantly expanded through quarterly releases. IMG is provided by the DOE-Joint Genome Institute (JGI) and is available from http://img.jgi.doe.gov.

INTRODUCTION

With about 20% of the reported genome projects worldwide, DOE-JGI is one of the main production centers of genome sequence data (1). IMG serves as a community resource for comparative analysis and annotation of all publicly available genomes from all three domains of life, in a uniquely integrated context.

Starting with version 2.0 released in December 2006, IMG has employed NCBI's RefSeq (2) as its main source of publicly available genomes. Through regular updates, IMG's data content has grown from a total of **296** genomes in its first version released in March 2005, to a total of **2,878** genomes in the version released in September 2007. New archaeal and bacterial genomes are added to IMG on a quarterly basis: IMG 2.3 (Sep 2007) has 729 bacterial and 46 archaeal genomes,. An increasing number of eukaryotic genomes, viruses (including phages) and plasmids have been also added to IMG in order to increase its genomic context for comparative analysis: IMG 2.3 has 50 eukaryotic genomes, 1,661 viruses, and 402 plasmids that did not come from a specific microbial genome sequencing project.

IMG's analytical tools have been gradually generalized and enhanced in terms of their usability, analysis flow, and performance. These tools allow users to focus on a subset of genes, genomes, and functions of interest, and conduct analysis using summary tables, graphical viewers, and various methods for comparing genes, pathways and functions across genomes.

DATA CONTENT AND CURATION

Genomes are identified in IMG via their taxonomic lineage (domain, phylum, class, order, family, genus, species, strain). For every genome, IMG incorporates its primary genome sequence information recorded in RefSeq including its organization into scaffolds and/or contigs, together with computationally predicted protein-coding sequences (CDSs) and some RNA-coding genes. IMG employs RefSeq's gene identifiers to link to other NCBI resources, such as Entrez Gene (3), and in order to establish gene based correlations with other microbial genome systems, such as Microbes Online (4).

Functional annotation of genes in IMG involves: (a) protein product assignment, (b) protein family and domain characterization, (c) IMG term assignment and (d) MyIMG protein function assignment. Protein product assignments are available from RefSeq and typically consist of the function prediction provided by sequence genome centres. Protein family and domain characterization involves

associating genes with various functional roles as defined in different controlled vocabularies, such as Enzyme Nomenclature (5), COG clusters (6), Pfam (7), TIGRfam (8), InterPro (9), Kegg Ortholog (KO) terms (10) and Gene Ontology (GO) terms (11). Genes are associated with COGs and Pfams based on RPS-BLAST (Reverse Position Specific BLAST) computation and NCBI's Conserved Domain Database (CDD) (12). EC numbers are computed using PRIAM (13), as a complement to the (often sparse) native EC numbers collected via RefSeq. UniProt (14) is used to associate genes with additional annotations, such as InterPro, TIGRfam, and GO terms, while KEGG is used to establish KO term associations. RNA gene models are synchronised with Rfam (15). Functional roles are further defined by their association with functional classifications including COG functional categories (6), TIGR role categories (8) and the KEGG pathway collection (10).

In order to address problems with the inconsistencies of the protein product assignments as well as with the current functional classifications (16), genes are further annotated in IMG using a native collection of generic (protein cluster-independent) functional roles called *IMG terms* that are further defined by their association with generic (organism-independent) functional hierarchies, called *IMG pathways*. IMG terms and pathways are currently specified by domain experts at DOE-JGI, as part of the process of annotating specific genomes of interest. IMG terms are subsequently propagated throughout the system. Finally, the users can add their own functional annotations which are captured under their user name as MyIMG annotations, as described below.

Homologs are computed as unidirectional hits with an E-value of 10^{-2} or better, with IMG providing support for filtering homolog lists by percent identity, bit score, and more stringent E-values, as well as with a variety of metadata such as phenotype, habitat, etc. In addition, CRISPR repeats (17), signal peptides using SignalP (18) and transmembrane helices using TMHMM (19) are computed, and potentially missing data from the original RefSeq data files (such as various RNAs) are added.

DATA ANALYSIS

Genome data analysis in IMG consists of operations involving genomes, genes, and functions which can first be ***selected*** and then ***explored*** individually. Genomes can be also ***compared*** in terms of various statistics, gene content, function capabilities, and sequence conservation.

Data Selection Tools

In order to perform comparative analysis in IMG, genomes, genes or functions are first selected using browsers or search tools. **Browsers** are provided for selecting genomes and functions, organized as alphabetical lists or hierarchically (e.g., based on phylogenetic tree for genomes). **Keyword search** tools allow identifying genomes, genes, and functions of interest using a variety of keyword filters. Genomes can be also selected using a search tool which allows specifying conditions involving phenotype, habitat, disease, and relevance metadata fields, while genes can be also selected using BLAST search tools against various datasets. The genomes that result from search operations are displayed as a list from which they can be selected and saved in order to reduce the genome context for further analysis. In a similar manner, the genes and functions that result from search operations are displayed as lists from which genes and functions can be selected for inclusion into the "**Gene Cart**" and "**Function Cart**", respectively.

Individual genomes can be explored using the "**Organism Details**" page which includes information on the organism together with various genome statistics of interest, such as the number of genes that are associated with KEGG, COG, Pfam, InterPro or enzyme information. For each genome one can also examine the associated list of scaffolds and contigs using the "**Chromosome Viewer**", or can generate circular chromosomal maps on which a variety of data can be projected.

Individual genes can be analyzed using the "**Gene Details**" page which includes Gene Information, Protein Information, and Pathway Information tables, evidence for functional prediction, COG, Pfam, and pre-computed homologs. A gene can be examined in the context of its location on the chromosome using the "**Chromosome Viewer**".

Individual functional groups, such as COG categories, can be further explored using summary pages, such as the "**COG Category Details**" page which lists the COGs of a given category and the number of organisms that have genes belonging to each COG, where the "organism counts" are linked to a list of organisms and their associated "gene counts".

Comparative Analysis Tools

Comparative analysis of genomes is provided in IMG through a number of tools that allow genomes to be compared in terms of various statistics, gene content, function capabilities, and sequence conservation.

"**Genome Statistics**" provides statistics across the genomes that have been previously selected and saved as discussed above. The display can be configured by including a variety of genome attributes, such as GC content, number of protein coding genes, and various functional annotations.

Genomes can be compared in terms of gene content using the "**Phylogenetic Profiler**" tool which allows to define a *profile* for the genes of the query genome, say archaeal genome *Thermoplasma volcanium GSS1 (T. volcanium)* in terms of presence or absence of homologs in any other genomes. In the example shown in pane (1) of Figure 1, the tool is used to find *T. volcanium* genes that have no homologs in *Thermoplasma acidophilum DSM 1728 (T. acidophilum)*. Similarity cutoffs can be used to fine-tune the selection. The list of genes with the specified profile are then provided as a selectable list as shown in pane (2) of Figure 1. The "**Phylogenetic Profiler**" tool can be used, for example for finding *unique* genes in the query genome with respect to other genomes of interest. In the example shown in Figure 1, 241 genes are found to be unique in *T. volcanium* with respect to *T. acidophilum.*

Genomes can be compared in terms of functional capabilities using the "**Abundance Profile Search**" tool which allows defining a *profile* for functions (COGs, Pfams) in a query genome in terms of their abundance compared to other related genomes.

In the example shown in pane (3) of Figure 1, this tool is used to find COGs that are more abundant in *T. volcanium* than in *T. acidophilum*. Some of the COG representatives found in *T. volcanium* (e.g. COG 1552) have no match in *T. acidophilum,* which may be of evolutionary significance or explained by the fact that the genes were missed by the original annotation. For each genome, a link to the list of genes associated with individual functions allows examining gene details.

The functional capabilities of genomes can be also compared using a number of additional functional profile tools. First, functions of interest, such as protein families, enzymes, IMG terms, are included into the "Function Cart", as illustrated in pane (5) of Figure 1. For these functions a profile across genomes can be computed, with the results displayed in a tabular format, as illustrated in pane (6) of Figure 1, with each column displaying the profile of a specific function across the genomes. The example in pane (6) of Figure 1 shows the profiles of several COGs of the *Signal transduction mechanisms* COG category across the *T. volcanium* and *T. acidophilum* genomes. Each cell in the profile result table displays the count (*abundance*) of genes in an organism and contains a link to the associated list of genes. Colours are used to represent visually gene abundance, whereby white, bisque and yellow represent gene counts of 0, 1-4, and over 4 respectively. The genes associated with a specific function can be saved using the "**Gene Cart**" and further examined using various tools, such as gene neighborhood analysis and multiple sequence alignment tools. For example, the "**Gene Ortholog Neighborhoods**" tool can be used to examine genes of *T. acidophilum* associated with a specific function (e.g., COG0467) together with its *T. volcanium* ortholog and their respective chromosome neighbourhoods, as shown in pane (7) of Figure 1.

Another functional profile tool, the "**Abundance Profile Viewer**", provides an ***overview*** of the relative abundance of protein families (COGs and Pfams) and functional families (Enzymes) across selected genomes, with abundance of protein/functional families displayed as a heat map. Note that the "**Function Cart**" in IMG provides users with the opportunity to define their own "pathways" and functional categories, assembled from individual COGs, Pfams, or Enzymes. Such user-defined "pathways" can be then employed in analysis of genomes and/or physiological traits that are poorly characterized by the traditional pathway databases, such as KEGG.

Comparative analysis of genes includes gene neighbourhood analysis, phylogenetic occurrence profile analysis and multiple sequence alignment, which can be applied to genes collected into the "**Gene Cart**".

Finally, DNA conservation can be explored for a number of organisms in IMG using the VISTA comparative genome analysis tools (20). Selecting an organism from a predefined list invokes the VISTA browser that can be then used for examining conservation.

User Annotations

IMG users can enter their own functional annotations using "**MyIMG"** tools, as illustrated in Figure 2. In this example, a gene of *Pyrococcus furiosus* is associated with product name *NADH oxidase*, as shown in pane (1) of Figure 2, and as recorded in GenBank and RefSeq. Based on a recent study

(21), it has been determined that the function for this gene is *NADPH:sulfur oxidoreductase*, and an expert review of the best homologs of this gene indicated that this product name also may be confidently applied to the top three homologs, as shown in pane (2) of Figure 2. The product name and several other gene annotations, such as the associated EC number, can be changed using "**MyIMG Annotation"** tool, as illustrated in pane (3) of Figure 2. User annotations are stored in IMG and can be reviewed at any time using the same tool. This tool also allows importing user annotations from user files (e.g. from excel files) into IMG or exporting user annotations in IMG to user files.

FUTURE PLANS

IMG continues to be extended in terms of data content through quarterly updates, whereby it aims at continuously increasing the number of genomes integrated in the system from public and local resources, following the principle that the value of genome analysis increases with the number of genomes available as a context for comparative analysis.

Future versions of IMG will focus on further improving the quality of gene models and functional annotations. We plan to expand the native IMG term controlled vocabulary and IMG pathway classification, jointly with annotation of IMG genomes using these terms and pathways. We also plan to provide extensive corroboration of annotations from other public microbial genome data resources, by including into IMG annotations based on TIGR genome properties (8) and MetaCyc (22). New data types such as results from microarray and proteomic experiments, as well as information on transcriptional regulatory binding sites will be also included into IMG.

IMG's analytical tools will continue to be extended in order to address two main challenges. First, as IMG's content expands, improved viewers will be developed in order to facilitate the exploration of a rapidly increasing number of genomes, genes, and annotations. Additional tools and viewers for exploring the power of gene context (i.e. fusions and gene neighbourhood) are also under current development. Since the comparative analysis context provided by IMG helps detect gene model and annotation errors, user annotation tools will be further extended based on requirements and feedback from the user community.

ACKNOWLEDGEMENTS

We thank, Philip Hugenholtz, Anu Padki, Kristen Taylor, Alla Lapidus, and Paul Richardson for their contribution to the development and maintenance of IMG. The work of JGI's production, cloning, sequencing, assembly, finishing and annotation teams is an essential prerequisite for IMG. Chris Oehmen of the Computational Biology and Bioinformatics group at the Pacific Northwest National Laboratory provided invaluable help in carrying out the large scale gene similarity computations for IMG 2.0. Eddy Rubin and James Bristow provided, support, advice and encouragement throughout this project. The work presented in this paper was supported by the Director, Office of Science, Office of Biological and Environmental Research, Life Sciences Division, U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

REFERENCES

1. Liolios, K., Tavernarakis, N., Hugenholtz, P., and Kyrpides, N. (2006) The genomes online database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Research* **34**, D332-D334.
2. Pruitt, K.D., Tatusova, T., Maglott, D.R. (2007) NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts, and proteins. *Nucleic Acid Research* **35**: D61-D65.
3. Maglott, D.R., Ostell, J., Pruitt, K.D., Tatusova, T. (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acid Research* **35**: D26-D31.
4. Alm, E.J., Huang, K.H., Price, M.N., Koche, R.P., Keller, K., Dubchak, I.L., Arkin, A.P. (2005) The Microbes Online web site for comparative genomics. *Genome Research* **15**(7): 1015-1022.
5. Bairoch A. (2000). The ENZYME database in 2000. *Nucleic Acids Research* **28**, 304-305.
6. Tatusov, R.L., Koonin, E.V., and Lipman, D.J., A. (1997) Genomic Perspective on Protein Families, *Science*, **278**, 631-637.
7. Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., *et al.* (2004) The Pfam Protein Families Database. *Nucleic Acids Research* **32**, D138-D141.

8. Selengut, J.D., Haft, D.H., Davidsen, T., Ganapathy, A., Gwinn-Giglio, M., Nelson, W.C., Richter, A.R., White O. (2007) TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes *Nucleic Acids Research* 35, D260-D264.
9. Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bradley, P., Bork, P., Bucher, P., Cerutti, L., et al. (2005) InterPro, Progress and Status in 2005. *Nucleic Acids Research* **33**, D201-D205.
10. Kanehisa, M., Goto, S., Kawashima, S. Okuno, Y., and Hattori, M. (2004) The KEGG Resource for Deciphering the Genome. *Nucleic Acids Research* **32**, D277-D280.
11. Gene Ontology Consortium. (2004) The Gene Ontology Database and Informatics Resource**.** *Nucleic Acids Research* **32**, 258-261.
12. Marchler-Bauer, A., Panchenko, A.R., Shoemaker, B.A., Thiessen, P.A., Geer, L.Y., Bryant, S.H. (2002) CDD: A Database of Conserved Domain Alignments with Links to Domain Three-Dimensional Structure. *Nucleic Acids Research* **30** (1), 281-283.
13. Claudel-Renard, C., Chevalet, C., Faraut, T., Daniel Kahn, D. (2003) Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Research* **31** (22), 6633-6639.
14. The UniProt Consortium. (2007) The universal protein resource (UniProt). *Nucleic Acids Research* **35**, D193-D197.
15. Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., Eddy, S.R. (2003) Rfam: an RNA family database. *Nucleic Acids Research* **31 (1)**, 439-441.
16. Ivanova, N.N., Anderson I., Lykidis A., Mavrommatis K., Mikhailova, N., et al. (2007) Metabolic Reconstruction of Microbial Genomes and Microbial Community Metagenomes. Technical Report 62292, Lawrence Berkeley National Laboratory.
17. Bland, C., Ramsey, T.L., Sabree, F., Lowe, M., Brown, K., Kyrpides, N.C., Hugenholtz, P. (2007) CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics*, **8**: 209.
18. Emanuelsson, O., Brunak, S., von Heijne, G., Nielsen, H. (2007) Locating proteins in the cell using TargetP, SignalP, and related tools. *Nature Protocols* **2**, 953-971.
19. Moller, S., Croning, M.D.R., Apweiler, R. Evaluation of methods for the prediction of membrane spanning regions. (2001) *Bioinformatics*, **17**(7), 646-653.
20. Frazer K.A, Pachter, L., Poliakov, A., Rubin, E.M., Dubchak, I. (2004) VISTA: Computational Tools for Comparative Genomics. *Nucleic Acids Research* **32**, W273-W279.
21. Schut, G.J., Bridger, S.L., Adams, M.W. (2007) Insights into the metabolism of elemental sulfur by the hyperthermophilic archaeon pyrococcus furiosus: characterization of a coenzyme a-dependent NAD(P)H Sulfur Oxidoreductase*. Journal of Bacteriology***.**
22. Caspi, R., Foerster, H., Fulcher, C.A., Hopkinson, R., *et al.* (2006) MetaCyc: a multiorganism database of metabolic pathways and enzymes, *Nucleic Acids Research* **34**, D511-D516.

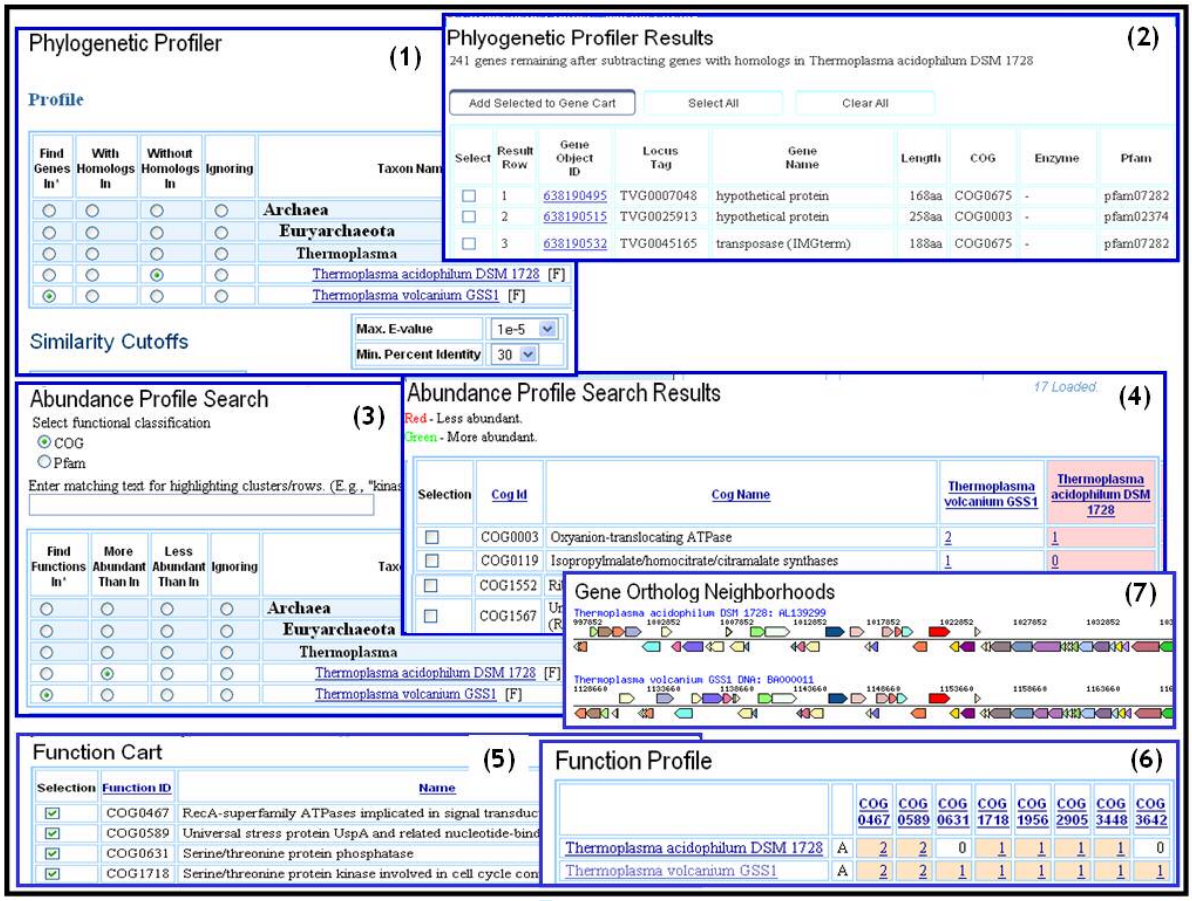

Figure 1. Profile Search and Functional Profile Tools. The "**Phylogenetic Profiler**" tool is used to find genes in the archaeal genome *Thermoplasma volcanium GSS1 (T. volcanium)* that have no homologs in a phylogenetically closely related genome, *Thermoplasma acidophilum DSM 1728 (T. acidophilum)* (1). Similarity cutoffs are used to fine-tune the selection. The list of *T. volcanium* genes that satisfy the specified condition are provided as a list (2) from which genes can be selected and added to a Gene Cart. The "**Abundance Profile Search**" tool is used to find COGs that are more abundant in *T. volcanium* than in *T. acidophilum* (3). The COGs that satisfy the specified condition are provided as a list (4) from which COGs can be selected and added to the Function Cart. For COGs of the *Signal transduction mechanisms* in the Function Cart, the "**Function Profile**" tool allows computing a profile across the *T. volcanium* and *T. acidophilum* genomes (5). The functional profile result is displayed as a matrix, with each cell containing a link to the associated list of genes and displaying the count (*abundance*) of genes in this list. Colours are used to represent visually gene abundance, whereby white, bisque and yellow represent gene counts of 0, 1-4, and over 4 respectively (6). The *T. volcanium* and *T. acidophilum* genes associated with COG0467 are further examined with the "**Gene Ortholog Neighborhoods**" tool (7): the compared genes are coloured red, and their respective neighbourhoods appear above and below a single line showing the genes reading in one direction on top and those reading in the opposite direction on the bottom; genes associated with the same colour indicate association with the same COG group.

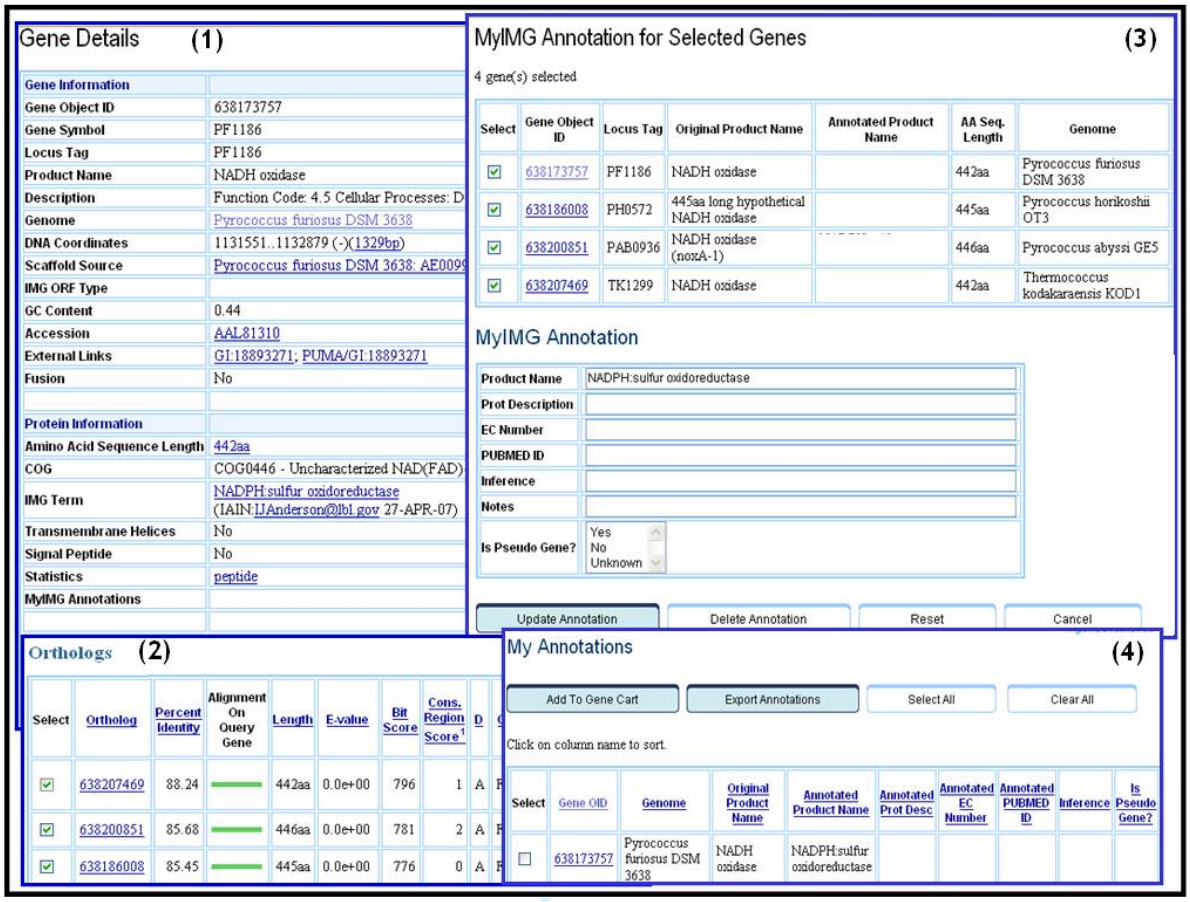

Figure 2. MyIMG User Annotations. A gene of *Pyrococcus furiosus* (with the object identifier "638173757") is associated with product name *NADH oxidase*, as displayed by its "Gene Details" page (1), and as recorded in GenBank and RefSeq, which can be viewed by following the appropriate External Links. For gene "638173757", one can display its list of orthologs, which can be accessed in the Homologs section of its "Gene Details" page (2). The top three orthologs of gene "638173757" can be selected and saved together with the gene into the Gene Cart, where the product name is changed to *NADPH:sulfur oxidoreductas* with the "**MyIMG Annotation**" tool (3). Other annotations (e.g., EC number) can be also modified. User annotations can be subsequently reviewed, imported into IMG from a user file, or exported to a user file using "My Annotations" page (4).